

公共部门绩效评估中的博弈行为： 分类框架、识别策略与研究进路

李佳源*

【摘要】政府绩效评估已成为公共管理研究的重要主题，但相当部分的文献仍然受到工具理性话语的主导，侧重于评估体系的构建和对评估价值的规范讨论。这些研究忽视了评估的“黑暗面”（Grizzle, 2002），即绩效评估引入公共部门后诱发了评估对象的策略性回应及其他一系列非预期后果。基于对多学科实证研究的梳理和整合，论文聚焦发生于公共部门情境下的绩效评估博弈行为，尝试归纳这类行为的一般模式以及近年来文献在这一主题研究上取得的方法上的突破。在此基础上，笔者指出现有研究的三个主要缺陷和可能的研究进路：采用非介入性的方法测量博弈行为的普遍性、构建中层解释框架，以及从组织行为等相邻学科中获取概念和分析工具。

【关键词】绩效评估 博弈行为^① 公共部门

【中图分类号】D63

【文献标识码】A

【文章编号】1674-2486（2021）03-176-20

一、引言

政府绩效评估是西方新公共管理改革中的重要举措，其逻辑起点是组织管理当中普遍存在的委托方和代理方之间的信息不对称以及由此产生的道德风险。为激励代理人，委托方将组织目标分解为具体、可测量的指标，然后按照目标的重要程度赋予相应的权重，并将评估结果应用于对代理人的奖惩（吴建南等，2009）。在工具理性话语的包装下，政府绩效评估与问责、透明、创新等话语关联起来，并逐渐成为一种制度化的迷思（Modell, 2004；Townley, 1997）：没有采用绩效评估的部门被认为管理落后，并可能导致组织合法性遭到质疑。因此，

* 李佳源，华南理工大学公共管理学院副教授。

① 在本文中，博弈行为特指评估对象为谋求相对有利的评估结果而采取的不正当手段。在相关外文文献中，主要的表述包括 strategic behavior、strategic response、gaming behavior 等。遵循这一表述惯例，本文交替使用策略行为和博弈行为来指代这类行为。

一个普遍的观察结论是，绩效评估成为一个迅速扩张的“产业”，现代组织面临着名目繁多的评估（Espeland et al., 2007; Espeland & Vannebo, 2007）。例如，俄罗斯颁布的第 825 号总统令要求俄联邦地区政府采集和发布多达 285 个绩效指标的相关数据（Kalgın, 2016）。英国布莱尔政府引入超过 300 个绩效指标对公共服务的绩效进行评价（Hood, 2006）。一些学者将量化评估向社会各个领域渗透的趋势形象地称为“审计社会”（audit society）（Power, 1997）。

随着绩效评估的扩散，针对这项管理工具实际影响的研究（即对评估的评估）成为一个重要的主题。一系列实证研究表明，绩效评估的推行在一定程度上提升了公共部门的效率（Propper et al., 2010），能有效激励地方政府推动经济的发展（Li & Zhou, 2005），显著改善公共服务质量（Dee & Wyckoff, 2015）并提升公共政策制定的理性化程度（Heinrich, 1999）。在新近发表的一项元分析中，围绕绩效评估是否能提升组织绩效这一核心问题，Gerrish（2016）对 49 份实证研究中报告的 2188 个效应量进行综合，作者发现评估体系与公共部门绩效的提升之间的相关性系数为 0.03（ $p < 0.05$ ），这证实了引入绩效评估可以提升公共部门的绩效水平。但这项研究所揭示的二者之间的弱相关性也表明，绩效评估的实施并没有如倡导者所期待的那样带来显著的效果（Osborne & Gaebler, 1993; Wholey & Hatry, 1992）。

事实上，绩效评估的价值，尤其是评估可能产生的负效应，一直是富有争议的话题。美国评估学报（*American Journal of Evaluation*）曾刊登了 4 篇针锋相对的评论文章（Bernstein, 1999; Perrin, 1998, 1999; Winston, 1999）。Perrin（1998, 1999）基于自己在公共部门的工作经历，对绩效评估的实践价值提出了质疑。他认为使用绩效指标来评价项目效果的做法存在“内在的错误和局限”。他列举了绩效评估的 8 个悖论来阐述这一管理实践的逻辑存在根本性的错误。例如，评估导致评估对象将注意力放在了数字生产上，而这些行为本身对绩效的提升并没有产生实际的意义。De Lancer Julnes（2006）同样指出了绩效评估面临的一系列挑战，特别是绩效信息并没有被应用于政策制定的过程。类似地，基于对两次绩效预算改革（GPRA 和 PART）的观察，美国国家行政科学院院士 Beryl Radin 指出了绩效评估的前提预设存在的问题（Radin, 2009）。例如，绩效评估的倡导者假定信息是客观中立的，绩效评估能明确项目效果产生的因果链条，政策目标是清晰的，以及公共部门的绩效是可量化的。然而，这一系列假设并不符合真实世界中的公共政策过程。值得注意的是，尽管侧重点不同，这些质疑均指向了绩效评估诱发的一系列非预期后果，尤其是代理人的策略性回应，这些行为构成了本文的主要关切点。

必须强调的是，尽管本文着力于从国际视野分析绩效评估的影响，但主题具有明显的本土关怀。自 20 世纪 80 年代以来，中国各级地方政府逐渐建立起一套量化的、强调实绩的绩效考核体系，进而对地方官员推动辖区经济和社会

发展形成有效的激励（荣敬本等，1998；王汉生、王一鸽，2009；周黎安，2008）。但这种“锦标赛式”的考核制度设计也诱发了一系列的负效应，其中最为突出的是考核中的绩效博弈行为，即评估对象对绩效考核进行不正当干预，以确保在绩效评估中的优势。自2017年以来，国内陆续有省市“自曝家丑”，承认经济数据造假，并主动“挤水分”，这些问题反映了绩效博弈的严重程度。显然，绩效博弈行为的发生对政府科学决策、官员的合理问责与激励以及政府纵向层级的治理构成巨大的挑战。

鉴于这类行为的危害性，本文尝试通过系统的文献综述回答以下两个主要问题。首先，政府绩效博弈行为的发生是否具有一般性的模式？鉴于博弈行为的隐秘性，许多相关研究都是选取某个特定的领域来对博弈行为进行深入分析，以便揭示博弈行为的存在及其发生机理。但这类研究设计容易导致研究发现的“碎片化”，许多实证研究停留在基于具体个案的经验叙事层次。诚如周雪光和练宏（2012）在评价中国政府研究时指出，政府行为的研究工作正进入一个收益递减阶段，即经验观察在不断积累，但却没有显著的知识增长。这个中肯的评价也适用于当前有关绩效博弈行为的研究。尽管我们获得了有关地方政府如何进行绩效博弈的经验观察，但尚未从中发掘出具有一般性、普遍性的知识贡献。对既有研究进行梳理、抽象和整合，并且尝试构建一个类型学，将是本文的目标之一。

其次，绩效博弈行为是一种敏感的组织行为，如何对这类行为的普遍程度进行测量构成了这一主题系统研究的主要方法挑战。不可否认，一些关于公共部门绩效博弈的知识仍局限于坊间传闻或日常生活中的直观性资料。近年来，方法上的创新为这一主题的研究注入了活力。一些学者采用非介入性的研究方法，对博弈的普遍性进行了系统的测量和建模，深化了我们对诱发博弈行为的组织和情境因素的理解。本文将致力于对这部分文献进行梳理和评价，并归纳测量博弈普遍程度的方法路径。

总之，本文希望在系统的文献综述基础上，对公共部门绩效博弈研究的分析工具、研究逻辑和方法设计进行归纳和评述，并反思这一主题研究的局限，提出若干有前景的研究方向。

二、核心概念的界定及文献检索方法

在介绍文献检索方法之前需厘清英文文献中有关绩效博弈的不同学术表达。在文献中，学者们经常把博弈（gaming）、策略行为（strategic behavior）、非预期后果（unintended consequences）、目标替代（goal displacement）和负功能（dysfunctional effect）等概念等同使用。笔者认为，目标替代指的是在考核的压力下，绩效评估作为组织激励的一种手段异化为目的本身，因此这些学术表达

指向的是博弈行为的后果，而非行为本身。一些学者如 Smith (1995) 使用“非预期后果”来描述引入绩效评估之后产生的事与愿违的后果。本文认为，博弈只是“非预期后果”的一种类型。部分学者倾向于使用 *perverse effects* 或 *dysfunctional effects* 来概念化实证观察。但这类标签隐含价值预设，即假定博弈是恶的。事实上，一些学者指出，在某些情形中，博弈行为尽管表面上违背了评估的初衷，但从结果上看并没有产生负面效果 (Kerpershoek et al., 2016)。综合考虑，本文使用博弈 (gaming) 这一概念来指代任何旨在干预绩效评估、且导致报告的业绩和实际业绩之间的背离的组织或个体行为。

为了系统地对绩效博弈这一主题的实证研究进行回溯、梳理和评述，本文按照循证管理学 (evidence-based management) 的文献检索程序定位尽可能多的相关研究 (Hoon, 2013; Rousseau et al., 2008)。为了确保进入综述的文献符合分析的需要，本研究确定以下 5 个纳入标准：

- (1) 只纳入以绩效评估为分析重点的实证研究。
- (2) 只纳入报告了某种形式的绩效博弈行为的实证研究。
- (3) 结合文献综述的目的，本文只分析发生在公共部门的绩效博弈行为。
- (4) 遵循系统文献综述的规范做法，纳入综述的研究必须符合事先确定的研究质量标准 (Shamseer et al., 2015)。例如，研究目标是否清晰？研究设计与研究问题是否匹配？研究是否采用了相关的措施来应对研究问题的敏感性？作者是否清晰地报告了研究发现？

(5) 纳入综述的研究必须发表于 1990—2016 年，且用英文写作。本文以 1990 年为检索起点是因为一系列有关公共部门绩效博弈的开创性研究发表于 20 世纪 90 年代初期 (Bouckaert & Balk, 1991; Smith, 1995)，这些研究影响了后续这一主题实证研究的发展轨迹。将著作发表的语言限定为英语，是为了最大限度拓展文献综述的国际视野，避免文献综述的结论囿于特定社会文化情境。

本文的检索词由绩效评估和博弈的关键词组合而成。其中，绩效评估在英文文献中的表达主要包括：performance system, performance management, performance measurement, performance indicator, performance standard, performance data, performance information, target regime。博弈相关的检索词组包括：cheat, falsity, game, manipulate, strategic behavior, strategic response, perverse effect, dysfunctional effect, unintended consequence。如表 1 所示，检索的数据库包括了社会科学研究常用的数据库。笔者根据每个数据库的检索规则来编写检索命令。以 SSCI 数据库为例，检索命令如下：

TS = (“performance system * ” OR “performance management” OR “performance measurement” OR “performance indicator * ” OR “performance standard * ” OR “performance data” OR “performance information” OR “target regime * ”) AND (cheat * OR falsif * OR gam * OR manipul * OR “strategic behavior \$ r * ” OR “strategic response * ”

OR “perverse effect *” OR “dysfunctional effect *” OR “unintended consequence *”).

表 1 文献检索记录

数据库	检索平台	检索时间段	检索结果
SSCI, A & HCI	Web of Knowledge	1990—2016	320 条文献记录
Asian Business Database, Periodicals Archive Online, ProQuest Dissertations and Theses A & I, Research Library	ProQuest	1990—2016	285 条文献记录
ERIC, MEDLINE, Psychology and Behavioral Science Collection, PsycINFO, eBook Collection, Communication & Mass Media Complete, Business Source Complete, Academic Search Complete	EBSCO	1990—2016	593 条文献记录
Scopus	Scopus	1990—2016	1493 条文献记录

笔者将检索得到的文献记录导入文献管理软件 Endnote X7 中, 得到总共 2735 条记录 (去重之后剩下 2078 条)。通过阅读标题和摘要, 并对照纳入标准进行筛选, 笔者初步确定其中 206 篇文献可以被纳入综述当中。通过对这些研究进一步研读, 确定最终纳入文献综述的研究。此外, 为了定位尚未正式发表的工作论文, 笔者利用类似的关键词在 SSRN^① 进行检索, 进一步获取与绩效博弈相关的文献。在此基础上, 笔者在公共管理 SSCI 刊物上检索与绩效博弈相关的实证研究, 这些刊物包括: *Journal of Public Administration Research and Theory*, *Public Administration Review*, *Public Administration*, *Public Management Review*, *Journal of Policy Analysis and Management*, *Governance*。鉴于聚焦中国情境的绩效博弈研究日益增长, 笔者在中国研究领域的英文刊物进行文献检索。这些刊物包括 *China Journal*, *China Quarterly*, *China Review*, *Journal of Contemporary China*, *Modern China*。

三、公共部门绩效评估中的博弈行为：一个分类框架

近年来, 海内外公共管理学界开始对绩效博弈行为进行系统研究, 逐步揭开了量化评估和博弈行为发生之间的联系 (Hood, 2007; Radnor, 2008)。对公

^① SSRN 全称为 Social Science Research Network, 该数据库致力于社会科学和人文科学领域研究的传播。研究人员可以将自己的工作论文上传至数据库中, 供读者免费下载和阅读, 并获得评论。根据最新的统计, SSRN 电子数据库收录超过 70 万篇论文, 研究成果覆盖 30 个学科。

共政策的效果进行量化评价，并以结果来激励公职人员是西方工业化国家新公共管理改革的主要举措（Kettl, 1997；Light, 1998）。但不少实证研究表明，这种做法在一定程度上诱发了公职人员的策略行为。随着实证研究的积累，一些学者开始尝试构建绩效博弈行为的类型学。Hood（2006）认为博弈可以分为三种类型。第一种策略行为被称为“棘轮效应”，指的是被考核者为了避免来年更重的任务，一般倾向于将产出维持在目标值附近。第二种策略行为被称为“门槛效应”，指的是考核对象一旦完成指标即停止进一步提升绩效的努力。第三种类型是对绩效数据的直接操纵。Gao（2015）提出了基于中国情境的博弈行为类型学。她认为，如果博弈导致报告的绩效和真实绩效之间的背离，那么这类行为属于恶性博弈（pernicious gaming）。在良性博弈（benign gaming）中，尽管存在策略性回应，但评估对象实际上完成了考核所要求的任务。Hood的类型学中的棘轮效应和门槛效应均属于这种类型。在恶性博弈中，评估对象尽管在数据上完成了考核目标，但实际业绩并没有达到考核的要求。

应该指出，类型化是提炼绩效博弈行为一般模式的有益尝试。但实际上，这些分类框架存在一定的局限。例如，博弈类型无法涵盖实证研究所揭示的诸多策略性回应；一些分类标准涉及明显的价值判断，因此对策略行为的分类也依赖于研究者的价值取向。通过对既有实证研究的梳理，本文认为绩效评估中博弈行为可以按照发生的节点来进行界定和区分，即发生在绩效生产和绩效验收两个阶段的策略行为。绩效生产的策略行为可以进一步细分为三类：曲解指标定义、选择性关注和干扰指标目标值的设定；绩效验收阶段的策略行为主要包括绩效数据篡改和扭曲抽样。

（一）绩效生产阶段的策略行为

绩效生产指的是作为代理方的评估对象如何完成委托方交付的工作，这些任务通常以可量化的指标来定义。文献梳理发现，在这一阶段可能出现三种主要的策略行为，即曲解指标定义、选择性关注和干扰指标目标值的设定。

1. 曲解指标定义

指标的定义是绩效评估的关键环节。尽管在制定评估方案时，委托方会对绩效目标做出明确的界定，但指标的定义存在不可避免的模糊性，在某些情况下需要依赖考核对象结合具体情境进行判断。这使得指标定义容易成为策略行为的风险点。文献梳理表明，在公共卫生和犯罪打击等专业性较强的领域，曲解指标定义成为一种较为普遍的博弈行为。

公共卫生医疗体系是新公共管理改革的关键政策领域。医护人员面临着名目繁多的考核指标，且评估的结果直接影响个人的业绩和医院的利益（Kelman & Friedman, 2009）。尽管一些实证研究提供了有力的证据表明绩效评估的应用提升了医疗质量（Bloom et al., 2015；Mason et al., 2012），但考核导致的非预

期后果也引起了学者们的关注。Hood 及其合作者的一系列研究揭示了量化考核的负效应 (Bevan & Hood, 2006; Hood, 2006, 2012)。例如, 为了提高公立医疗的效率, 英国国民医疗服务系统 (National Health Service) 规定急诊病人在英格兰公立医院的等待时间不得超过 4 个小时。大规模的就诊数据分析表明这个预期目标得到了实现 (Mason et al., 2012)。但在考核的压力下, 一些医院为了确保能达到这个目标, 让病人在救护车中等待, 直到确定可以在这个规定的时间内接诊才允许他们进入医院 (Bevan & Hood, 2006)。类似地, 为了保证病患的及时就医, 卫生部门为信托机构设定了不同急救类型的响应时间要求 (从接到明确求救电话到救护设备抵达现场), 对于没有达到预定标准的信托, 将面临包括声誉受损乃至行政主管被撤的一系列惩处措施。一系列内部审计结果表明, 一些医院为了达到考核目标采取了不正当的手段。例如, 类别 A 是指那些可能危及生命的求救, 因此属于紧急程度最高的急救类别, 卫生部门规定医院信托 75% 的类别 A 求助电话必须在 8 分钟内响应, 但一些信托机构利用类别定义存在的模糊性, 将一些原本应归为类别 A 的求助降格为类别 B 或 C, 从而降低响应时间的标准。

绩效评估在公共卫生领域的另一个主要应用是将医护人员的工资待遇与评估结果挂钩, 但学界对于绩效工资激励效果存有争议 (Doran et al., 2008; Eijkenaar et al., 2013; Li et al., 2014)。聚焦英国国民医疗服务的绩效工资体系, Gravelle et al. (2010) 发现, 推行绩效工资制度扭曲了医护人员的激励。2004 年引入的质量和结果评估框架要求英国全科诊所 (General Practice) 报告 146 个质量指标上的业绩, 其中 65 个指标是通过计算在某个疾病类型中诊所治疗的病人数量 (分子) 和患有该类疾病的总人数 (分母) 的比例来衡量的, 以此评估诊所在该具体疾病类型上的绩效 (总共涉及 11 种疾病)。当该比值达到设定的阈值上限, 诊所可以在这个指标上获得最高的分值。

尽管采用比值 (ratio) 的初衷是为了激励诊所收治更多的病患 (即计算比重的分子部分), 但为了获得更多的奖金, 诊所可以通过减少分母的方法来达到同样的目的。由于分母是诊所报告的患有特定类型疾病的病人数量减去基于正当理由不纳入计算的病患数量, 一个可能的干预策略就是增加免除纳入指标计算的病患数量。借助苏格兰地区 932 个诊所的基础数据, Gravelle 及其合作者发现, 2004—2005 财务年度指标数据低于阈值上限的诊所, 相比于指标数据高于阈值上限的诊所, 前者在下一财务年度免除纳入计算的病患比例明显更高。这说明, 诊所可能通过操纵指标的分母部分来提升绩效。

犯罪数字的生产过程以及犯罪数据的失真一直是社会学研究的重要主题 (Black, 1970; Gove, et al., 1985; Price, 1966; Skogan, 1974; Seidman & Couzens, 1974)。破案率是评价警察部门绩效的重要指标。尽管法律明确规定了哪些行为构成犯罪, 但对行为的定性在一定程度上依赖于警方的判断 (Patrick,

2009)。追求高破案率在一定程度上导致立案数据不实的问题。其中，“有案不立”便是一种典型的策略行为。长期服役于英格兰警队的帕特里克的研究发现(Patrick, 2009)，为提升破案率这一关键指标，警察拒绝登记受害人向警方报告的犯罪行为，或将发案时间和地点相似的独立案件合并登记为1宗案件（做小分母）。类似的策略行为也存在于中国地方警务绩效的考核。例如，Li（2015）发现，派出所为完成上级下达的“治安拘留数”，人为降低拘留标准，将一些轻微的违法行为纳入拘留的范围。同样聚焦犯罪打击领域的数字生产，Xu（2018）的研究揭示了在“两升一降”的考核压力下（即破案率、逮捕数上升，立案数下降），一些公安机关采取包括拒绝立案在内的一系列办法来追求表面上的高绩效，这些策略行为严重影响了绩效数据的真实性。

从上述的例子可以看出，这类策略行为的共性都是行动者利用绩效标准定义的模糊性干扰考核结果。

2. 选择性关注

组织注意力是一种稀缺资源，当组织面临多重任务要求时，量化考核容易导致选择性关注，即考核对象根据服务对象和工作任务对评估结果的影响来分配注意力。这类策略行为具体表现为“筛选”服务对象或忽视那些没有被纳入考核的绩效维度。以就业培训为例，Heinrich（2007）对美国劳工投资法案的分析发现，尽管这是一个普惠性的福利项目，但由于绩效标准是根据培训参与者结束培训时的就业情况来确定，一些就业培训机构为了获取绩效奖励，将一些就业前景不佳的申请者（如少数族裔）排除在项目之外。

筛选服务对象的行为也存在于公共卫生领域。在医疗领域引入绩效评估的一个重要做法是公布医院的绩效信息。这一举措的初衷是为了解决病患和服务供给方之间的信息不对称问题，让病人做出更理性的就医选择。但是，由于治疗失败将直接影响医院在关键绩效指标上的表现，医院可能拒绝接收风险较高的病患。一些实证研究证实了医疗服务提供者存在规避高风险病患的倾向。例如，Dranove et al.，（2003）评估了发布医院冠状动脉搭桥术的死亡率产生的非预期后果。作者借助双重差分法分析发现，引入了强制性的信息公开机制之后，纽约州和宾州接受搭桥手术的病人的病情严重程度，相比于没有引入类似机制的州明显下降，这说明医院存在选择病人的可能。

在日趋强调结果和量化评估的教育领域，教师在考核的压力下出现选择性关注也是较为普遍的现象。美国联邦政府在2002年通过了《不让一个孩子掉队法案》（*No Child Left Behind Act*），引入以标准化测试中学生通过率为关键指标的考核体系。研究发现，法案生效之后，学校的一个策略性回应是将主要精力放在考试及格线附近的学生。原因是相比于成绩特别优秀或特别差的学生，这部分学生的表现将直接影响学校的整体通过率以及排名（Reback, 2008）。

在环保领域，中国政府自“十一五”规划对各省的二氧化硫、化学需氧量

两类污染物的排放削减情况进行考核。尽管减排考核促使地方政府重视属地的环境保护,但这种做法也在一定程度上诱发了类似的策略行为。利用31个省份2001—2010年的面板数据,Liang和Langbein(2015)发现,考核显著减少了二氧化硫的排放量,但对于化学需氧量的削减并未产生明显的影响。作者认为,化学需氧量反映的是水体的污染,而二氧化硫反映的是空气的污染程度,后者对环境的影响更加直观,因此更可能引起地方政府的重视。此外,两位研究人员发现,对于那些未纳入考核的污染指标,其排放情况并未显著改善。这些证据表明,环保量化考核扭曲了地方环保部门的注意力分配——那些没有被纳入考核或难以直接观察的绩效维度被相对忽视。

上述的文献综述表明,尽管指标量化可以对代理方的工作进展和绩效进行监测,但由于信息不对称的问题,代理方如何完成这些工作难以体现在指标数据上。评估本身是一个“去情境化”(de-contextualization)的过程,考核对象在完成指标任务的过程中,非常容易出现测量什么就重视什么(What's measured is what matters.)的问题。

3. 干扰指标目标值的设定

考核目标值的高低直接关系到完成考核任务的难易以及通过考核的可能性,因此目标值的设定也是博弈高发的环节之一。在实践过程中,考核目标通常依据评估对象过往的绩效状况而定,因此容易出现“鞭打快牛”的情况,即绩效越高的部门可能承担更重的绩效要求。因此,为避免更重的考核任务,评估对象倾向于压低产出。Bevan & Hood(2006)将这类行为称为棘轮效应(ratchet effect)。

韩国中央政府推行绩效预算改革产生的非预期后果便是一个典型的例子。为提升公共部门的绩效,韩国政府仿效美国的《政府业绩与成果法案》(Government Performance and Results Act),引入绩效预算体系。2005年,中央政府引入了预算项目自评计划(Self-Assessment of Budgetary Programmes, SABP),明确建立起绩效和财政拨款之间的关联。这个项目被用于评估支出机构超过1/3的项目。从2005—2010年,总计有2920个项目接受评估,涉及财政资金约1320亿英镑。按照规定,项目绩效评估结果划分为五档:非常有效、有效、一般、差、很差。若等级为“差”的,该项目下一年度的预算将被削减10%。在其博士论文中,Shin(2013)针对中央政府官员的访谈发现,为了避免预算受到影响,部门上报的绩效目标偏于保守,以确保能通过考核。

除此以外,干扰指标目标值的做法可能涉及组织之间的共谋。Li(2015)的研究发现,在犯罪打击领域,个别基层派出所之间存在指标“交易”行为。例如,一个派出所在逮捕数这个指标上超额完成,但在行政拘留数这一指标上的完成情况不理想,而另外一个派出所的指标完成情况恰好相反。在这种情况下,两个派出所可以按照一定比例在两个不同的考核指标上进行交易。这种行

为实质上属于干扰预期目标的设定，其结果是交易双方的“共赢”：没有完成既定目标的基层所队能通过考核，而超额完成任务的单位可以避免来年更重的任务。

（二）绩效验收阶段的策略行为

绩效验收特指委托方对绩效的完成数量和质量按照事先确定的标准进行评估。文献梳理的结果表明，这个过程容易发生两类策略行为：一类是直接对评估的结果数据进行篡改，另外一类是在抽样阶段对数据进行干扰。

1. 绩效数据篡改

篡改绩效数据是最为典型的策略性回应。在教育领域，学生的学业成绩已成为评价教师和学校的关键业绩指标。一方面，部分实验研究表明，引入激励机制明显提升了教学效果和学生的学业表现（Behrman et al., 2015; Muralidharan & Sundararaman, 2011; Victor Lavy, 2002）。但另一方面，将学生测试成绩用于评价教师也会诱发数据造假的问题。在这一研究主题上，Jacob 和 Levitt（2003）的论文已成为广为引用的经典研究。作者借助测试成绩的异常波动和同班学生测试答案的可疑分布规律来推断教师修改试卷答案的可能性。利用来自芝加哥公立学校的数据，作者保守估计每年至少有 4% ~ 5% 的小学班级存在试卷答案被篡改的情况。

近年来，现场考察（inspection）逐渐成为一种重要的评估方式，评估专家现场的评价直接影响教育机构的绩效。但这也诱发了各种“迎评”策略：学校和教师通过对文本的精心制作和场景的布置来进行“印象管理”。Perryman（2009）的民族志研究颇具代表性。作者对英格兰一所地处内城的学校进行了历时 3 年的参与式观察，并参加了英国教育标准局（Office for Standards in Education）对该校的现场督学。作者考察的这个学校由于没有通过标准局组织的上一轮督查而面临巨大的压力。在现有评估框架下，督学团若发现学校的绩效没有明显提升，那么教育标准局可以对学校采取包括关停在内的手段。她观察到，为了应对考核，教师和管理人员花了大量的时间来准备评估文件、工作计划和课程大纲。由于督学团要求对教师和学生进行面谈，校方组织多场排练以帮助教员更好地应对各种可能的提问。为了不影响督查结果，一些学生甚至被藏匿或转移。这些迎评策略实质上带有很强的表演性，督学团所看到的远非教育机构的真实运作情况。

过去十几年间，基于中国情境的绩效博弈行为研究呈现增长的趋势。作为改革开放以来衡量地方政绩的核心指标之一，中国经济数据在多大程度上受到干预成为一个被热烈讨论的话题。Rawski（2001）通过 GDP 的增长和能源消耗之间的反差对中国 GDP 数据的真实性提出质疑。作者对 1997—2000 年的经济增长数据进行分析后发现，3 年间全国 GDP 累计增长 24.7%，而同期全国能源消耗下降 12.8%。作者认为，经济增长是建立在能源消耗的基础上，GDP 快速增

长而能耗下降只有两种解释,一种是单位 GDP 能耗剧烈下降,这显然不符合当时的实际,另外一种解释就是数据失真。结合中国的官员选拔制度,作者认为 GDP 数据注水的可能性更大。沿袭类似的分析思路,徐康宁、陈丰龙和刘修岩(2015)提出以夜间灯光数据作为检验经济数据真实性的替代测量,这是由于许多研究证实夜间灯光亮度与经济增长之间高度相关(Chen & Nordhaus, 2011; Henderson et al., 2012)。对 1992—2012 年的数据进行分析后发现,基于灯光数据测算的全国实际 GDP 增长率比同时期官方数据平均低 1.02 个百分点,其中东部、中西部比官方数据分别低 1.52 个百分点和 1.81 个百分点。据此,作者推断地方政府存在夸大经济数据的可能性。

近年来,非经济政策领域的策略行为也引起了研究人员的关注。以环保领域为例,数据真实性一直是环境经济学研究备受关注的议题(Andrews, 2008; Brombal, 2017; Stoerk, 2016; Zheng et al., 2014)。Chen 等(2012)较早对环境保护领域数据造假进行系统的研究。他们借助两个方法来检验数据是否受到操纵。首先是利用相对独立的数据对地方政府发布的空气质量数据进行校验。第一个指标是能见度。作者认为该指标没有用于考核地方政府,因此受到干预的可能性较小。另一个指标是美国国家航空航天局发布的气溶胶光学厚度^①(AOD),该数据同样具有较高的独立性。作者的具体假设是,如果地方政府发布空气污染指数(API)数据在临界点附近受到操纵,那么 API 与能见度、AOD 这两个空气污染数据的替代测量之间的相关性会显著减弱。其次,他们借助断点回归的方法来判断博弈是否存在。其中的逻辑是,由于“蓝天数”的关键指标之一——API 存在明确的阈值,即 API 小于 100 才达标,如果数据被人为干预,那么 API 在临界点之下的密度将低于 API 在临界点之上的密度。借助中国 37 个主要城市 2000—2009 年的每日 API、能见度和 AOD 等数据,作者发现官方数据在临界点附近出现明显的间断,而且这种跳跃在引入“蓝天数”考核指标之后更加明显,这说明 API 值在一定程度上受到了人为的干预。

2. 干扰抽样

绩效验收实际上是一个抽样的过程,即从数据总体中抽取一定数量的样本进行评估,进而推断总体的绩效状况。推断的准确性取决于样本的代表性。为谋求有利的考核结果,作为考核对象的代理人有动机对采样过程进行干扰。以教育领域为例,在引入《不让一个孩子掉队法案》之后,由于特殊教育的学生的测试成绩不会影响学校排名,校方可能将考试成绩较差的学生归入特殊教育的类别。Jacob(2005)的研究发现,引入问责政策之后,特殊教育的学生比例增加了 8%,并且排名越靠后的学校的增加幅度越大。这个发现与 Bohte 和 Meier

^① 大气气溶胶(Aerosol Optical Depth, 缩写为 AOD)通常是指悬浮在大气中的固态或液态微粒,是衡量大气混浊度的重要指标。

(2000) 针对德克萨斯州公立学校的发现基本一致。另外一种策略行为是将学业表现不佳的学生留级，避免他们升入三年级、六年级和八年级这三个需要参加标准化学业测试的年级。作者发现，引入法案之后，一、二年级留级的学生比例显著增加（增加 2.3 个百分点），其中，排名靠后的学校的留级学生增幅最大。类似的情况出现在四、五、七年级这三个年级中。

干扰抽样的策略行为还可能出现在采样的时点上（timing）。在就业培训领域，为了准确衡量培训机构的绩效，美国劳工部要求培训机构在项目参与者培训终止后 90 天内报告其就业状况。Courty 和 Marschke（1997）围绕美国《劳工投资法案》（*Workforce Investment Act*）的研究表明，由于培训机构在报告时点的选择上有一定的自由裁量权，培训机构根据当年的整体业绩情况采取不同的报告策略。当整体绩效不佳，机构倾向于推迟报告未就业者的就业状况，尽可能确保参与者能够在报告截止时间前获得工作。当培训机构的业绩超过了预定的绩效标准时，项目机构立即终止那些没有接受培训且尚未就业的项目参与者。显然，在机构业已完成绩效目标的情况下，报告不利结果并不会影响机构的绩效（reporting bad outcomes in good times only）。

总之，绩效博弈行为理论上可能发生在绩效评估的诸多环节。通过对实证研究的梳理和归纳，本文将这类行为按照发生的时间节点划分为两类，即绩效生产阶段和绩效验收阶段的策略行为。其中，绩效生产阶段的策略行为包括了曲解指标定义、选择性关注和干扰指标目标值设定；绩效验收阶段的策略行为主要包括绩效数据篡改和干扰抽样两种类型。

四、如何测量绩效博弈行为

上述文献综述表明，相当部分的研究尝试通过与知情人进行深入访谈，获得关于策略行为的证据。但作为一个敏感的话题，直接询问依赖访谈对象的配合，因此研究结论的可信度难免会受到质疑。近年来，参与式观察提供了近距离研究博弈行为的机会，特别是一些具有一线工作经验的公职人员的研究，更是打开了这一主题研究的黑箱，极大地丰富了学界对一些具体政策领域博弈行为的认识。但必须指出的是，揭示博弈行为发生机制的前提是对博弈行为的普遍程度进行有效度的测量。博弈属于敏感的组织行为，因此，如何对博弈行为进行识别和测量是这个主题研究的重点和难点。本文在文献检索中定位到一些初步的尝试，这些研究借助系统的定量方法对博弈行为的普遍程度进行建模，揭示了博弈行为发生的可能机制。基于对这部分文献的分析，本文将既有研究测量博弈程度的方法分为两类。

第一种方法是通过相对独立的数据来验证官方报告的绩效数据的真实性，并利用二者之间的背离估计博弈行为的普遍程度。Chen 等（2012）的研究是以

空气污染的替代测量和绩效数据的相关性来分析后者是否被操纵。Wallace (2016) 和徐康宁 (2015) 等人的研究在此基础上尝试量化博弈的普遍程度。这两份研究把发电量/用电量和地区夜间灯光亮度作为对地区经济活动活跃程度的替代测量, 然后将其与官方报告的经济增长数据进行比较, 二者之间的差距被认为是策略行为所致。其中的逻辑是: 在一般情况下, 经济增长和发电量/用电量或夜间灯光亮度之间是高度相关的, 二者间的背离可能是数据操纵导致。

Wallace (2016) 的研究显然更进一步。他巧妙地利用了发电量/用电量的增长率和官方报告的 GDP 增长率的差异作为对经济数据造假程度的估计, 证实了时间节点 (timing) 对于解释组织博弈行为发生的重要性。基于 2000—2009 年的省级面板数据, 他发现, 在省委书记或省长出现人事变动时, GDP 增长率和用电量/发电量增长率之间的背离增大约 1 个百分点。这个发现在控制了一系列重要变量后仍然稳健。这表明, 经济数据在政治敏感时期更可能被操纵。必须指出的是, 尽管这种估计方法具有较大的创新性, 但正如徐康宁等 (2015: 28) 所指出的, 客观数据测算的结果和官方数据的差异可能是人为干扰所致, 也可能是源自技术性误差, 因此也就不能将二者的差异等同于数据造假。要揭示博弈行为发生的内在机制, 必须对这类行为发生的普遍程度进行直接测量 (笔者在讨论部分提出新的研究思路)。

采用相对独立的数据对官方报告的绩效数据进行校验也被应用于计划生育领域。以 1990 年、1995 年和 2000 年实施的全国人口普查数据, 以及国家计生委 1988 年、1992 年和 1997 年的全国生育调查为参照数据, Serrato 等 (2019) 分析了地级市政府报告的人口计生绩效数据的真实性。他们发现, 地方政府报告的人口自然增长率为 7.7‰, 而使用全国人口普查数据估计的增长率为 8.3‰。作者将基于普查数据计算得到的人口出生率和地方政府报告的出生率比较, 二者的差异被视为数据操纵所致。一个重要的发现是, 地方政府报告的出生率和实际出生率之间的差异在审计年份的前一年明显缩小。这也反证了二者的差距来源并非测量误差, 而可能是地方政府对数据的操纵。这是因为如果不存在人为的干预, 我们不太可能观察到二者之间的背离在审计年份会减小。

测量博弈的第二种分析思路是借助数据分布规律的异常来反推博弈的发生和普遍性。在 Jacob 和 Levitt (2003) 的研究中, 作者构建了 4 个指标来衡量学生试卷答案的异常分布, 以此间接推断舞弊的普遍性。例如, 同一个班级的学生的答案在一些不可能的情况下出现高度一致 (答错容易的题目但答对了较难的题目)。这些指标允许作者对考试舞弊的普遍程度进行估计, 并构建统计模型来探索诱发舞弊的可能因素。例如, 作者的分析发现, 将学生在标准化考试中的成绩和学校的招生资格等一系列激励措施关联后, 班级舞弊的可能性上升, 尽管二者的相关性没有通过显著度检验。Bohte 和 Meier (2000) 对德克萨斯州公立学校的博弈行为研究也采用了类似的方法。作者认为校方在考试通过率的

压力下，可能将学习成绩较差的学生排除在考试之外，从而人为提升学区的整体通过率。基于这样的逻辑，作者利用非正常免试的学生比例来测量博弈行为的普遍程度，并分析了包括绩效压力、资源的充裕程度、任务难度等因素与非正常免试比例之间的统计关系^①。

必须指出的是，尽管这些方法比较巧妙，但局限于某个具体政策领域。相比较而言，数位分布规律和断点回归是两种具有应用前景的测量方法。一些研究人员借助本福德法则（Benford's Law）对数据的真实性进行校验。根据该法则，一组随机发生的数字的第一位数字在1-9之间分布的频率并非完全一致，具体而言，越小的数字，出现的概率越大（Nigrini, 2012）。例如，首位数为1的可能性约为30%，而数字9出现的频率为4.6%。这个规律存在于许多不同的数据集，例如街道的门牌号、人口数量、死亡率及河流的长度等。根据这个法则，如果数据被操纵，那么数字的首位数分布将不符合本福德法则。经济学家把这个方法用于检验经济和财务数据的真实性（Holz, 2014；Michalski & Stoltz, 2012）。Stoerk（2016）将这个研究方法运用于分析中国空气质量数据的“统计腐败”。作者对三组数据进行分析：第一组数据是北京环保局公布的每天空气质量数据（2007—2013年），第二组数据是美国大使馆公布的北京市空气PM_{2.5}数据（2008—2013年），第三组数据是从全球气溶胶监测网获取的北京市AOD的数据（2008—2013年）。从数据来源推测，作者认为第一组数据存在造假的可能性较大，而第二和第三组数据相对独立，因此数据点首位数的分布应符合本福德法则。分析的结果与作者的假设高度吻合，作者据此判断北京的空气质量存在人为操纵的可能。

近年来，经济学家开始重视采用断点回归的方法来识别博弈证据。这个分析方法的实质也是利用数据分布的异常来推断数据是否遭到篡改。Fisman和Wang（2017）对中国安全生产事故死亡人数的真实性分析便是采用断点回归统计方法。借助不同类型事故死亡人数的面板数据（2005—2012年），作者发现，在引入“一票否决”问责制之后，安全生产事故死亡率超过上级规定的上限的可能性显著降低。这说明绩效考核能有效减少安全生产事故的数量。但与此同时，借助断点回归的方法，作者发现，省级政府的安全生产事故的死亡人数在目标值这一临界点出现间断，这表明数据可能遭到人为的调整。这类研究方法的共性是在选取问责制度中设定的关键目标值作为临界点进行分析。基本逻辑是，如果不存在操纵行为，那么数据在临界点两边的分布是连续的；相反，如果存在操纵行为，那么绩效数据的分布将出现“跳跃”。由于很多绩效评估均存在明

① 以绩效压力为例，作者采用两个关键指标来衡量学校的绩效压力。首先是学区的考试通过率，其次是学区1994—1995年之间整体通过率的变化趋势。通过率低和考试成绩下降的学区，绩效考核的压力更大，更加可能采用不正当的手段对数据进行人为的调整。分析结果表明，绩效差距与考核博弈的测量之间缺乏显著的统计相关性。

确的目标值（即临界点），随着公共部门绩效数据的公开，利用断点回归的方法来研究绩效博弈将是一种有前景的分析工具。

五、对文献的反思及可能的研究进路

以上的文献综述表明，绩效考核诱发的各类策略行为业已成为公共管理研究和中国研究中重要的领域。不同学科之间通过共享研究问题推动了这一主题的知识积累，丰富并加深了我们对地方政府绩效博弈行为发生过程的认识。但绩效博弈的实证研究仍存在若干局限。

首先，缺乏有效的研究工具分析绩效博弈行为的普遍程度。尽管公共管理学界已经意识到对绩效博弈行为进行学术研究的重要性，但由于其敏感性使得对这一组织现象进行实证研究的难度较大。诚如公共管理研究学者 Christopher Pollitt 指出，分析博弈的普遍性构成“最难，但同时也是最重要的问题”（Pollitt, 2013: 357）。传统的社会科学研究方法，例如问卷调查或访谈法都很难获得博弈实施者的真实回应，因此也无法对这些行为的普遍性进行有效度的测量。事实上，一些学者指出，我们对绩效评估体系中博弈和欺骗行为的普遍程度仍知之甚少（Bevan & Hood, 2006）。本文的文献梳理发现，近年来绩效博弈研究主题出现了一些方法上的创新，例如，Wallace（2016）和徐康宁等（2015）的研究通过独立数据来间接推断博弈行为的发生：借助与经济状况紧密相关、且相对真实的用电量或夜间灯光亮度来间接推断经济增长数据的真实性。尽管这类研究工具由于不依赖研究对象的自我报告，可信度比较高，但采用这类研究方法实质上研究的是策略行为的后果，无法估计行为本身的普遍程度。因此，对政府绩效博弈行为的实证研究急需方法上的创新，这将是今后研究的着力点之一。

在研究工具的选取上，笔者认为可以将社会科学中对敏感话题的分析方法延伸至博弈主题的研究。例如，研究者可以尝试借助列举实验（list experiment）对博弈行为的普遍性进行测量，最新的研究参见 Kroll 和 Vogel（2021）。列举实验是对传统抽样调查方法的改进，其创新之处在于将随机化的操作程序引入调查中，从而帮助研究者更严谨地推断敏感行为的普遍程度。许多研究表明，相比于直接询问的方法，列举实验对应答者的隐私有着更好的保护，因此参与者对敏感问题的回答更真实（Rosenfeld et al, 2016）。例如，Gonzalez-Ocantos et al.（2012）针对尼加拉瓜城市选举的调查发现，在问卷调查中 2.39% 的受访者报告自己曾有过被贿选经历，而通过列举实验估计的比例达到 24.34%。绩效策略行为是敏感的话题，直接询问非常容易获得不真实的回应。在新近的一份基于 54 项实验研究的元分析中，Li 和 Van den Noortgate（2019）发现，在估计敏感行为/态度的普遍性上，列举实验的有效性显著高于直接询问法。因此，列举实验

法为测量这类行为的普遍程度提供了一种可行的技术方案。研究人员可以利用这个方法对博弈行为的普遍性进行测量，并且围绕博弈行为的发生构建多层次的解释模型^①。

绩效博弈研究的第二个主要局限集中在博弈发生的解释上。既有研究对绩效博弈行为诱因的分析存在两个极端。一是基于具体个案获得的解释，但这些解释由于“就事论事”而缺乏推论性；另一极端是从制度层面对博弈行为发生的原因作出回答，但解释框架过于宏大而缺乏分析的力度（例如制度逻辑的视角），且难以支撑实证研究。在这一点上，罗伯特·默顿所强调的“中层理论”（mid-level theory）可以帮助我们在宏观和微观解释之间求得平衡。所谓“中层理论”是指介于无所不包的抽象综合性理论和具体经验性观察二者之间的理论层次。具体到绩效博弈行为研究，中层理论模型指的是超越碎片化的经验观察基础上形成的具有普遍意义的解释。将绩效博弈研究提升到理论化的层面可以避免实证观察的简单堆砌，实现从对特殊经验的揭示迈向具有普遍意义的新认识，这是后续研究的努力方向。

笔者认为，对既有文献实施更为系统的整合分析有助于构建一个中观层面的解释框架。前面提到，尽管文献已对博弈行为提供了丰富的经验观察，而且既有的研究已揭示了一些重要的解释变量，主要包括官员晋升的节点、被考核者的自由裁量权、强调竞争的组织文化、绩效目标的可及性、绩效信息收集的独立性、完成考核所需资源的稀缺程度、过于强调自上而下的问责制，等等。尽管这些因素涉及了不同层面，但均是基于某个特定领域的案例研究，对既有研究的整合还有待进一步提升。研究者可以借助对相关文献进行系统的综合，实现在抽象层面对文献所报告的各种博弈行为进行比较和综合，为构建预测博弈行为发生的初步模型提供概念基础。

缺乏不同学科之间的交融和互惠，这是绩效博弈行为研究的第三个主要局限。笔者认为，对绩效博弈的研究可以尝试从组织研究和管理学研究中获取概念和分析工具的启发，从而为这个主题的研究注入更多的活力。本文的文献综述涉及公共管理、组织研究、教育学、经济学、环境保护和公共卫生等领域。尽管这些研究共同指向绩效评估过程中的一个基本悖论，但这些研究在概念、分析框架和研究方法上缺乏交融，具体表现在实证研究之间鲜有交叉互引。例如，管理学研究中的组织不端行为（unethical behavior）与博弈行为之间存在概念上的相关性，但绩效博弈研究者对这个概念缺乏关注。根据管理学的广义理解，不端行为是指组织成员任何违背普遍接受的（社会）道德规范的行为（Kish-Gephart et al., 2010）。因此，尽管组织不端行为是组织研究的视角，但绩效博弈行为实质上也属于组织不端行为，可以放到这一理论框架下进行研究。

① 该技术新近的发展已经允许研究者在不同分析层面上识别影响个体实施敏感行为的因素，具体应用可参见 Holbrook 和 Krosnick（2010）。

在一份新近发表在 *Public Administration Review* 的研究中, Bellé 和 Cantarelli (2017) 对 73 篇论文 137 个实验实施元分析, 并归纳出导致组织不端行为的机制, 包括社会影响、目标设定、时间压力、自我辩解、监督、损失规避 (loss aversion) 等因素。如何从这些一般性的理论中获取分析框架, 并在公共部门绩效评估的情境中进行实证检验, 可以成为今后研究的努力方向。

六、余论

美国社会心理学家 Campbell (1979: 85) 提出了著名的坎贝尔法则: “愈是使用量化的社会指标作为社会决策的依据, 就愈容易产生腐化的压力, 也就愈容易曲解和破坏原本想要监控的社会历程。” 这个悲观的观点不断被实证研究所证实。公共部门目标的模糊性以及目标之间的相互冲突导致了绩效博弈更容易发生。打破坎贝尔法则, 绩效博弈研究必须系统深入地研究博弈行为如何发生以及其产生的机理。本文认为这一领域的研究取得突破的关键点在于如何通过研究方法上的创新来揭示不同考核领域中博弈行为的普遍程度; 如何在理论层面进行跨学科的对话, 构建一个关于绩效博弈行为的类型、特征和诱因的中层模型。通过研究方法和理论框架的创新, 以产生指导实践的知识为目标, 这些洞见对构建博弈行为发生的风险监测与控制体系、改善地方政府绩效评估的努力将起到积极的作用。

参考文献

- 荣敬本、崔之元 (1998). 从压力型体制向民主合作体制的转变: 县乡两级政治体制改革. 北京: 中央编译出版社.
- Rong, J. B., & Cui, Z. Y. (1998). *Transformation from a Pressured System to a Democratic System of Cooperation*. Beijing: Central Compilation & Translation Press. (in Chinese)
- 涂子沛 (2014). 大数据: 正在到来的数据革命, 以及它如何改变政府、商业与我们的生活. 桂林: 广西师范大学出版社.
- Tu, Z. P. *The Big Data Revolution*. Guilin: Guangxi Normal University. (in Chinese)
- 王汉生、王一鸽 (2009). 目标管理责任制: 农村基层政权的实践逻辑. 社会学研究 2: 61-92+244.
- Wang, H. S., & Wang, Y. G. Target Management Responsibility System: The Practice Logic of Local Part-State in Rural China. *Sociological Studies*, 2: 61-92+244. (in Chinese)
- 吴建南、章磊、李贵宁 (2009). 地方政府绩效指标设计框架及其核心指标体系构建. 管理评论 11: 121-128.
- Wu, J. N., Zhang, L., & Li, G. N. Research on the Framework of Performance Indicator for Local Government and the Construction of Its Key Performance Indicators. *Management Review*, 11: 121-128. (in Chinese)
- 徐康宁、陈丰龙、刘修岩 (2015). 中国经济增长的真实性: 基于全球夜间灯光数据的检验. 经济研究 9: 17-29+57.
- Xu, K. N., Chen, F. L., & Liu, X. Y. The Truth of China Economic Growth: Evidence from Global Night-time Light Data. *Economic Research Journal*, 9: 17-29+57. (in Chinese)
- 周黎安 (2008). 转型中的地方政府: 官员激励与治理. 上海: 格致出版社.
- Zhou, L. A. (2008). *Local Government in Transition: Official Incentive and Governance*. Shanghai: Truth & Wisdom Press. (in Chinese)
- 周雪光、练宏 (2012). 中国政府的治理模式: 一个“控制权”理论. 社会学研究 5: 69-93+243.
- Zhou, X. G., & Lian, H. Modes of Governance in the Chinese Bureaucracy: A “Control Rights” Theory. *Sociological Studies*, 5: 69-93+243. (in Chinese)
- Andrews, S. Q. (2008). Inconsistencies in Air Quality Metrics: ‘Blue Sky’ days and PM₁₀ Concentrations in Beijing.

- Environmental Research Letters*, 3(3): 034009.
- Behrman, J. R., Parker, S. W., Todd, P. E., & Wolpin, K. I. (2015). Aligning Learning Incentives of Students and Teachers: Results from a Social Experiment in Mexican High Schools. *Journal of Political Economy*, 123(2): 325–364.
- Bellé, N., & Cantarelli, P. (2017). What Causes Unethical Behavior? A Meta-Analysis to Set an Agenda for Public Administration Research. *Public Administration Review*, 77(3): 327–339.
- Bernstein, D. J. (1999). Comments on Perrin's "Effective Use and Misuse of Performance Measurement". *The American Journal of Evaluation*, 20(1): 85–93.
- Bevan, G., & Hood, C. (2006). What's Measured Is What Matters: Targets and Gaming in the English Public Health Care System. *Public Administration*, 84(3): 517–538.
- Black, D. J. (1970). Production of Crime Rates. *American Sociological Review*, 35(4): 733–748.
- Bloom, N., Propper, C., Seiler, S., & Van Reenen, J. (2015). The Impact of Competition on Management Quality: Evidence from Public Hospitals. *The Review of Economic Studies*, 82(2): 457–489.
- Bohte, J., & Meier, K. J. (2000). Goal Displacement: Assessing the Motivation for Organizational Cheating. *Public Administration Review*, 60(2): 173–182.
- Bouckaert, G., & Balk, W. (1991). Public Productivity Measurement: Diseases and Cures. *Public Productivity & Management Review*, 15(2): 229–235.
- Brombal, D. (2017). Accuracy of Environmental Monitoring in China: Exploring the Influence of Institutional, Political and Ideological Factors. *Sustainability*, 9(3): 324.
- Campbell, D. T. (1979). Assessing the Impact of Planned Social Change. *Evaluation and Program Planning*, 2(1): 67–90.
- Chen, X., & Nordhaus, W. D. (2011). Using Luminosity Data as a Proxy for Economic Statistics. *Proceedings of the National Academy of Sciences*, 108(21): 8589–8594.
- Chen, Y., Jin, G. Z., Kumar, N., & Shi, G. (2012). Gaming in Air Pollution Data: Lessons from China. *BE Journal of Economic Analysis and Policy*, 12(3): 1–43.
- Courty, P., & Marschke, G. (1997). Measuring Government Performance: Lessons from a Federal Job-Training Program. *The American Economic Review*, 87(2): 383–388.
- Dranove, D., Kessler, D., McClellan, M., & Satterthwaite, M. (2003). Is More Information Better? The Effects of "Report Cards" on Health Care Providers. *Journal of Political Economy*, 111(3): 555–588.
- De Lancer Julnes, P. (2006). Performance Measurement: An Effective Tool for Government Accountability? The Debate Goes On. *Evaluation*, 12(2): 219–235.
- Dee, T. S., & Wyckoff, J. (2015). Incentives, Selection, and Teacher Performance: Evidence from IMPACT. *Journal of Policy Analysis and Management*, 34(2): 267–297.
- Doran, T., Fullwood, C., Reeves, D., Gravelle, H., & Roland, M. (2008). Exclusion of Patients from Pay-for-Performance Targets by English Physicians. *New England Journal of Medicine*, 359(3): 274–284.
- Eijkenaar, F., Emmert, M., Scheppach, M., & Schöffski, O. (2013). Effects of Pay for Performance in Health Care: A Systematic Review of Systematic Reviews. *Health Policy*, 110(2–3): 115–130.
- Espeland, Wendy N., & Sauder, M. (2007). Rankings and Reactivity: How Public Measures Recreate Social Worlds. *American Journal of Sociology*, 113(1): 1–40.
- Espeland, W. N., & Vannebo, B. I. (2007). Accountability, Quantification, and Law. *Annu. Rev. Law Soc. Sci.*, 3: 21–43.
- Fisman, R., & Wang, Y. (2017). The Distortionary Effects of Incentives in Government: Evidence from China's. *American Economic Journal: Applied Economics*, 9(2): 202–218.
- Gao, J. (2015). Pernicious Manipulation of Performance Measures in China's Cadre Evaluation System. *The China Quarterly*, 223: 618–637.
- Gerrish, E. (2016). The Impact of Performance Management on Performance in Public Organizations: A Meta-Analysis. *Public Administration Review*, 76(1): 48–66.
- Gonzalez-Ocantos, E., de Jonge, C. K., Melendez, C., Osorio, J., & Nickerson, D. W. (2012). Vote Buying and Social Desirability Bias: Experimental Evidence from Nicaragua. *American Journal of Political Science*, 56(1): 202–217.
- Gove, W. R., Hughes, M., & Geerken, M. (1985). Are Uniform Crime Reports a Valid Indicator of the Index Crimes? An Affirmative Answer with Minor Qualifications. *Criminology*, 23(3): 451–502.
- Gravelle, H., Sutton, M., & Ma, A. (2010). Doctor Behaviour under a Pay for Performance Contract: Treating, Cheating and Case Finding? *The Economic Journal*, 120(542): F129–F156.
- Grizzle, G. A. (2002). Performance Measurement and Dysfunction: The Dark Side of Quantifying Work. *Public Performance & Management Review*, 25(4): 363–369.

- Heinrich, C. J. (1999). Do Government Bureaucrats Make Effective Use of Performance Management Information? *Journal of Public Administration Research and Theory*, 9(3): 363–394.
- Heinrich, C. J. (2007). False or Fitting Recognition? The Use of High Performance Bonuses in Motivating Organizational Achievements. *Journal of Policy Analysis and Management*, 26(2): 281–304.
- Henderson, J. V., Storeygard, A., & Weil, D. N. (2012). Measuring Economic Growth from Outer Space. *The American Economic Review*, 102(2): 994–1028.
- Holbrook, A. L., & Krosnick, J. A. (2010). Social Desirability Bias in Voter Turnout Reports: Tests Using the Item Count Technique. *Public Opinion Quarterly*, 74(1): 37–67.
- Holz, C. A. (2014). The Quality of China's GDP Statistics. *China Economic Review*, 30: 309–338.
- Hood, C. (2006). Gaming in Targetworld: The Targets Approach to Managing British Public Services. *Public Administration Review*, 66(4): 515–521.
- Hood, C. (2007). Public Service Management by Numbers: Why Does It Vary? Where Has It Come from? What Are the Gaps and the Puzzles? *Public Money & Management*, 27(2): 95–102.
- Hood, C. (2012). Public Management by Numbers as a Performance-Enhancing Drug: Two Hypotheses. *Public Administration Review*, 72(s1): S85–S92.
- Hoon, C. (2013). Meta-Synthesis of Qualitative Case Studies: An Approach to Theory Building. *Organizational Research Methods*, 16(4): 522–556.
- Jacob, B. A. (2005). Accountability, Incentives and Behavior: The impact of High-stakes Testing in the Chicago Public Schools. *Journal of Public Economics*, 89(5–6): 761–796.
- Jacob, B. A., & Levitt, S. D. (2003). Rotten Apples: An Investigation of the Prevalence and Predictors of Teacher Cheating. *The Quarterly Journal of Economics*, 118(3): 843–877.
- Kalgin, A. (2016). Implementation of Performance Management in Regional Government in Russia: Evidence of Data Manipulation. *Public Management Review*, 18(1): 110–138.
- Kelman, S., & Friedman, J. N. (2009). Performance Improvement and Performance Dysfunction: An Empirical Examination of Distortionary Impacts of the Emergency Room Wait-Time Target in the English National Health Service. *Journal of Public Administration Research and Theory*, 19(4): 917–946.
- Kerpershoek, E., Groenleer, M., & de Bruijn, H. (2016). Unintended Responses to Performance Management in Dutch Hospital Care: Bringing Together the Managerial and Professional Perspectives. *Public Management Review*, 18(3): 417–436.
- Kettl, D. F. (1997). The Global Revolution in Public Management: Driving Themes, Missing Links. *Journal of Policy Analysis and Management*, 16(3): 446–462.
- Kish-Gephart, J. J., Harrison, D. A., & Treviño, L. K. (2010). Bad Apples, Bad Cases, and Bad Barrels: Meta-analytic Evidence about Sources of Unethical Decisions at Work. *Journal of Applied Psychology*, 95(1): 1–31.
- Kroll, A., & Vogel, D. (2021). “Why Public Employees Manipulate Performance Data: Prosocial Impact, Job Stress, and Red Tape. *International Public Management Journal*, 24(2): 164–182.
- Li, H., & Zhou, L. A. (2005). Political Turnover and Economic Performance: The Incentive Role of Personnel Control in China. *Journal of Public Economics*, 89(9–10): 1743–1762.
- Li, J. (2015). The Paradox of Performance Regimes: Strategic Responses to Target Regimes in Chinese Local Government. *Public Administration*, 93(4): 1152–1167.
- Li, J., Hurley, J., DeCicca, P., & Buckley, G. (2014). Physician Response to Pay for Performance: Evidence from a Natural Experiment. *Health Economics*, 23(8): 962–978.
- Li, J., & Van den Noortgate, W. (2019). A Meta-analysis of the Relative Effectiveness of the Item Count Technique Compared to Direct Questioning. *Sociological Methods & Research*, online.
- Liang, J., & Langbein, L. (2015). Performance Management, High-powered Incentives, and Environmental Policies in China. *International Public Management Journal*, 18(3): 346–385.
- Light, P. C. (1998). *The Tides of Reform: Making Government Work, 1945–1995*. New Haven, CT: Yale University Press.
- Mason, S., Weber, E. J., Coster, J., Freeman, J., & Locker, T. 2012. Time Patients Spend in the Emergency Department: England's 4-Hour Rule: A Case of Hitting the Target but Missing the Point? *Annals of Emergency Medicine*, 59(5): 341–349.
- Michalski, T., & Stoltz, G. (2012). Do Countries Falsify Economic Data Strategically? Some Evidence That They Might. *The Review of Economics and Statistics*, 95(2): 591–616.
- Modell, S. (2004). Performance Measurement Myths in the Public Sector: A Research Note. *Financial Accountability & Management*, 20(1): 39–55.
- Muralidharan, K., & Sundaraman, V. (2011). Teacher Performance Pay: Experimental Evidence from India. *Journal of Political Economy*, 119(1): 39–77.
- Nigrini, M. (2012). *Benford's Law: Applications for Forensic Accounting, Auditing, and Fraud Detection*. John Wiley & Sons.

- Osborne, D. , & Gaebler, T. (1993). *Reinventing Government: How the Entrepreneurial Spirit Is Transforming the Public Sector*. New York: Plume.
- Patrick, R. (2009). *Performance Management, Gaming and Police Practice: A Study of Changing Police Behaviour in England and Wales during the Era of New Public Management*. University of Birmingham.
- Perrin, B. (1998). Effective Use and Misuse of Performance Measurement. *The American Journal of Evaluation*, 19(3): 367–379.
- Perrin, B. (1999). Performance Measurement: Does the Reality Match the Rhetoric? A Rejoinder to Bernstein and Winston. *The American Journal of Evaluation*, 20(1): 101–111.
- Perryman, J. (2009). Inspection and the Fabrication of Professional and Performative Processes. *Journal of Education Policy*, 24(5): 611–631.
- Pollitt, C. (2013). The Logics of Performance Management. *Evaluation*, 19(4): 346–363.
- Power, M. (1997). *The Audit Society: Rituals of Verification*. Oxford: Oxford University Press.
- Price, J. E. (1966). A Test of the Accuracy of Crime Statistics. *Social Problems*, 14(2): 214–221.
- Propper, C. , Sutton, M. , Whitnall, C. , & Windmeijer, F. (2010). Incentives and Targets in Hospital Care: Evidence from a Natural Experiment. *Journal of Public Economics*, 94(3–4): 318–335.
- Radin, B. A. (2009). What Can We Expect from Performance Measurement Activities? *Journal of Policy Analysis and Management*, 28(3): 505–512.
- Radnor, Z. (2008). Muddled, Massaging, Maneuvering or Manipulated? A Typology of Organizational Gaming. *International Journal of Productivity and Performance Management*, 57(4): 316–328.
- Rawski, T. G. (2001). What Is Happening to China's GDP Statistics? *China Economic Review*, 12(4): 347–354.
- Reback, R. (2008). Teaching to the Rating: School Accountability and the Distribution of Student Achievement. *Journal of Public Economics*, 92(5–6): 1394–1415.
- Rosenfeld, B. , Imai, K. , & Shapiro, J. N. (2016). An Empirical Validation Study of Popular Survey Methodologies for Sensitive Questions. *American Journal of Political Science*, 60(3): 783–802.
- Rousseau, D. M. , Manning, J. , & Denyer, D. (2008). 11 Evidence in Management and Organizational Science: Assembling the Field's Full Weight of Scientific Knowledge Through Syntheses. *The Academy of Management Annals*, 2(1): 475–515.
- Seidman, D. , & Couzens, M. (1974). Getting the Crime Rate Down: Political Pressure and Crime Reporting. *Law & Society Review*, 8(3): 457–493.
- Serrato, J. C. S. , Wang, X. Y. , & Zhang, S. (2019). The Limits of Meritocracy: Screening Bureaucrats under Imperfect Verifiability. *Journal of Development Economics*, 140 (2019): 223–241.
- Shamseer, L. , Moher, D. , Clarke, M. , Ghersi, D. , Liberati, A. , Petticrew, M. , Shekelle, P. , & Stewart, L. A. (2015). Preferred Reporting Items for Systematic Review and Meta-analysis Protocols (PRISMA – P) 2015: Elaboration and Explanation. *BMJ*, 349.
- Shin, S. H. (2013). *Dysfunctional Consequences of the Korean Performance Budgeting System and Their Policy Implications*. University of Birmingham, Birmingham.
- Skogan, W. G. (1974). The Validity of Official Crime Statistics: An Empirical Investigation. *Social Science Quarterly*, 55(1): 25–38.
- Smith, P. (1995). On the Unintended Consequences of Publishing Performance Data in the Public Sector. *International Journal of Public Administration*, 18(2): 277–310.
- Stoerk, T. (2016). Statistical Corruption in Beijing's Air Quality Data Has Likely Ended in 2012. *Atmospheric Environment*, 127: 365–371.
- Townley, B. (1997). The Institutional Logic of Performance Appraisal. *Organization Studies*, 18(2): 261–285.
- Victor Lavy. (2002). Evaluating the Effect of Teachers' Group Performance Incentives on Pupil Achievement. *Journal of Political Economy*, 110(6): 1286–1317.
- Wallace, J. L. (2016). Juking the Stats? Authoritarian Information Problems in China. *British Journal of Political Science*, 46(1): 11–29.
- Wholey, J. S. , & Hatry, H. P. (1992). The Case for Performance Monitoring. *Public Administration Review*, 52(6): 604–610.
- Winston, J. A. (1999). Performance Indicators—Promises Unmet: A Response to Perrin. *The American Journal of Evaluation*, 20(1): 95–99.
- Xu, J. (2018). Legitimization Imperative: The Production of Crime Statistics in Guangzhou, China. *The British Journal of Criminology*, 58(1): 155–176.
- Zheng, S. , Kahn, M. E. , Sun, W. , & Luo, D. (2014). Incentives for China's Urban Mayors to Mitigate Pollution Externalities: The Role of the Central Government and Public Environmentalism. *Regional Science and Urban Economics*, 47: 61–71.

责任编辑：王秋石